(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 182 581 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
27.02.2002 Bulletin 2002/09

(51) Int Cl.[7]: **G06F 17/30**

(21) Application number: **00402311.5**

(22) Date of filing: **18.08.2000**

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(71) Applicant: **Exalead**
75019 Paris (FR)

(72) Inventors:
- **Bertin, Patrice**
 78800 Houilles (FR)
- **Bourdoncle, Francois**
 75014 Paris (FR)

(74) Representative: **Cabinet Hirsch**
34, Rue de Bassano
75008 Paris (FR)

(54) Searching tool and process for unified search using categories and keywords

(57) A database of entries, such as Web pages and sites is provided. The entries are at least partially mapped to a set of predetermined categories. The entries are also associated with keywords, for instance by automatic indexing of documents. In response to a query (1) into the database, a user is provided with a series of refinement strategies (13, 15, 17), in addition to search results (7). Refinement strategies comprise categories relevant for the search, selected among the set of predetermined categories. Refinement strategies also include keywords dynamically selected among keywords associated with the entries.

The user may easily navigate among the results to the query, and formulate new queries.

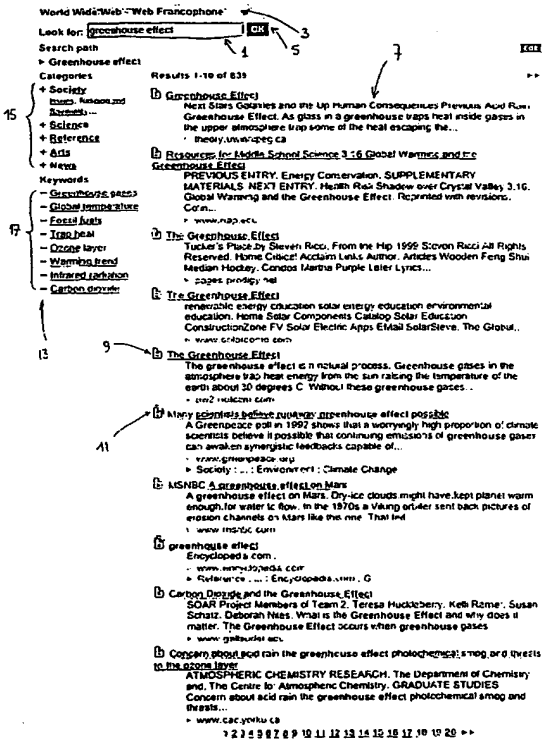


FIG. 1

Printed by Jouve, 75001 PARIS (FR)

EP 1 182 581 A1

BNSDOCID: <EP_____1182581A1_I_>

Description

[0001] The invention relates to the field of information retrieval, and more specifically to displaying results to a search query, as well as navigating in databases and inputting requests to databases. It particularly applies to searches on the Internet.

[0002] Throughout the present specification, the word "site" or "internet site" refers to a number of documents connected by links, with a given entry point. A directory is the result of indexing a number of sites or documents and of classifying these into categories; categories are therefore subsets of the directory, which are usually defined in a manual operation; such categories are often organised in a tree to facilitate navigation among categories; one may also use categories organised in a directed acyclic graph, that is a graph with a plurality of paths to the same category. A search engine is a tool for searching among documents, usually embodying automatic indexing of the documents.

[0003] A number of searching tools exist for searching and retrieving information on the Internet. Alta Vista Company proposed an Internet search site, with a request box where the user may input keywords for retrieving information. The language of the search may be restricted. A box is provided, allowing the user to select related searches; the related searches actually display phrases or sequences of words, which contain the current request as a substring. For instance, if the request inputted by the user reads /greenhouse effect/ (in the rest of this specification, the request will be marked by //), related searches could offer the following choices:

- the greenhouse effect,
- what is the greenhouse effect,
- enhanced greenhouse effect.

[0004] There is also proposed a search among site categories; this search is actually an independent category search in a separate database.

[0005] The results of the search are displayed to the user under the list of related searches. The results are displayed as a list of documents or sites.

[0006] Another Internet search site is proposed by Yahoo ! Inc. There is again provided a request box. Results of a search inputted to the request box are displayed in several sections. The first section displays the category matches, together with the path to the matches in the category tree, while the second section displays site matches. The third section displays web pages.

[0007] With the same example of /greenhouse effect/, the first category match is "global warming". The path to "global warming" in the category tree is Home > Society and Culture > Environment and Nature. There may be provided several paths to the same category; in the example of /greenhouse effect/, the category entitled "global warming" appears in five different paths. Selecting a category in the first section allows the user to access the contents of the category.

[0008] The second section displays site matches; matches are clustered according to their categories. The third section displays web pages, together with a summary and an address.

[0009] Google Inc also provides an Internet site for search among sites and categories. The results of a search contain an indication of the classification of sites and categories. When inputting the keywords for a search, some words may be excluded. Selecting the category search provides the user with a list of categories that may relate to the search; the contents of each category may later be accessed. In the example of the /greenhouse effect/ search, categories comprise Society / Issues / Environment / Climate Change.

[0010] A.V. Leouski and W. Bruce Croft , An Evaluation of Techniques for Clustering Search Results, CIIR Technical Report IR-76, National Center for Intelligent Information Retrieval, University of Masachussetts Amherst, MA, U.S.A., Spring 1996, compare classification methods from Information Retrieval and Machine Learning for clustering search results in a search engine. Apart from clustering techniques, this document discusses cluster description. A first method for describing a cluster consists in selecting a number of the most important terms from the documents comprised in the cluster, and in presenting them to the user. A second preferred method is to replace the important terms with important phrases - a phrase being defined as a sequence of one or more words. This document provides a solution to the problem of dynamically clustering documents retrieved from a database by a search engine.

[0011] US-A-5 463 773 discloses the building of a document classification tree by recursive optimisation of keyword selection function. There is provided retrieval means for extracting keywords when a document data is inputted, and outputting a classification for the document data, the classification being selected among the classification decision tree. For extracting keywords, this document suggests extracting keywords defined by word sequences. A learning process is suggested for building automatically a document classification tree on the basis of the extracted keywords.

[0012] US-A-5 924 090 proposes searching among documents, and mapping the keywords of the documents among static categories. Categories are therefore predefined in a manual process. The use of categories makes it possible to access documents comprised in the categories, that are mapped to the categories. In this document, a search engine provides the results of a query; the results are mapped onto the static categories, and relevant categories are displayed to the user as search folders. When a search folder is selected by the user, the documents comprised in the search folder - that is, the documents mapped onto the corresponding category - are displayed to the user. A series of search folders is displayed any time a search is carried out, the search folders being those static categories into

BNSDOCID: <EP_____1182581A1_I_>

which a number of documents retrieved were mapped.

[0013] US-A-5 963 965 discloses a method where relevant sets of phrases are automatically extracted from text-based documents in order to build an index for these documents. These phrases are then grouped together in clusters to form a plurality of maps which graphically describe hierarchical relationships between the clusters, and can be used to extract relevant portions of the documents in answer to the user selecting one of these clusters.

[0014] US-A-5 991 756 describes a method according to which search queries may be applied to a set of documents organised in a hierarchy of categories, and where the user is presented in response with a subset of these categories which contain the documents relevant to the query.

[0015] WO-A-98 49637 suggests organising results of a search into a set of most relevant categories. In response to a search, the search result list is processed to dynamically create a set of search result categories. Each of the search result categories is associated with a subset of the records within the search result list having common characteristics. Categories are then displayed as folders.

[0016] The prior art information retrieval methods and processes have a number of shortcomings. Fixed or static categories actually provide a representation of the world - a set of documents - at a given time point and for a given field of the art. They may need updating, or adapting to new types of documents, when and if the set of documents is completed by new documents, especially by documents in a new field of the art. While static categories may therefore represent accurately the expertise of the human being who defined them, they are in fact limited to this expertise. In addition, any set of categories is limited by the amount of human work needed for completing categories and mapping entries of the database to the categories.

[0017] Clusters formed of keywords may provide a dynamic vision of the word. However, they do not provide an easily browsable tool, and do not allow the user to navigate easily and freely among documents.

[0018] Category search is adapted to searching among sites. Keyword searches are more adapted to searching among separate textual documents.

[0019] Therefore, there is a need for an information retrieving process and tool, that enables a user to navigate not only among fixed categories, but also among keywords. The invention thus proposes a searching tool and process enabling its user to freely navigate among categories and keywords, in a friendly and transparent fashion. The invention combines the advantages of a set of human-made categories - notably expertise in a given field - together with the advantages of a keyword search - notably the ability to process and handle documents outside of said given field -. The invention provides a tool that is well adapted to searching among a database of sites and separate documents or pages.

[0020] More specifically, the invention provides a process for searching a database of entries, comprising the steps of:

- providing a database of entries, at least part of said entries being mapped to a set of categories, at least part of said entries being associated with keywords;
- in response to a query of a user, selecting categories among said set of categories according to the entries returned by said query ;
- dynamically selecting keywords associated to the entries returned by said query ; and
- displaying to the user said selected categories and said selected keywords.

[0021] In a particular embodiment of the process, the categories are organised in a tree or directed acyclic graph structure. A keyword may preferably be a sequence of words or a sequence of stemmed words.

[0022] Selected categories and keywords may be displayed similarly, or separately.

[0023] In a particular embodiment of the process, a new query is started when a user activates one of said displayed categories and keywords. This step of activating may comprise refining the said query of the user to the said category or keyword. This step of activating may also comprise excluding from the said query of the user a displayed category or keyword.

[0024] In another embodiment of the process, a list of entries returned by the query is displayed to the user. One may then display in said list a category to which at least an entry of said list is mapped. When the user selects this category in the list, the entries comprised in the selected category may be displayed. In this case the entries comprised in said category may be ranked before they are displayed.

[0025] Additionally, when categories are hierarchically organised, the step of displaying may comprise displaying categories of different hierarchical levels.

[0026] The invention also provides a searching tool, comprising a search server for receiving queries from users and transmitting results to users, a database of entries, at least part of said entries being mapped to a set of categories, at least part of said entries being associated with keywords; wherein the search server comprises

- means for searching the database and for selecting categories among said set of categories according to the entries returned by said query ;
- means for dynamically selecting keywords associated to the entries returned by said query;
- and wherein the results transmitted to the users comprise said selected categories and said selected keywords.

[0027] The search server is preferably a HTTP server. When the entries are textual entries, the database may

BNSDOCID: <EP_____1182581A1_I_>

comprise an inverted index, the categories being entries of said inverted index.

[0028] A search tool embodying the invention will now be described, by way of example only, and in reference to the attached drawings, where

- figure 1 is a display of a searching tool according to the invention ;
- figure 2 is another view of the display of the searching tool of figure 1, after the search is limited to a category ;
- figure 3 is another view of the display of the searching tool of figure 1, after the search is limited to a keyword;
- figure 4 is another view of the display of the searching tool of figure 2, after the search is limited to a keyword;
- figure 5 is another view of the display of the searching tool of figure 1, after the search is limited to a category of the list of results;
- figure 6 is a flowchart of a process according to the invention;
- figure 7 is a diagram of a searching tool according to the invention.

[0029] In response to a query, the invention suggests displaying to the user a series of refinement strategies; the refinement strategies comprise relevant categories selected among a set of static categories. In addition, the series comprises keywords dynamically obtained from the documents provided in response to the query. The invention allows the user to refine the search using predefined categories; in addition, displaying keywords allows the user to navigate more easily among the results of the query, without being limited to the fixed categories.

[0030] In the rest of the specification, the invention is disclosed in reference to its preferred embodiment; the database covers the World Wide Web, and comprises Internet sites as well as Web documents. The tool according to the invention allows the database to be searched thanks to a HTTP server.

[0031] More generally, the invention may apply to any database were entries are at least partially mapped to predetermined categories, and may be associated with keywords. Mapping is usually a manual operation, although it is possible to use any automatic process. Textual entries may easily be associated with keywords, e. g. by automatically indexing the entries and selecting keywords. In this case, automatic mapping to categories may be carried out based on keywords.

[0032] Figure 1 is a display of a searching tool according to the invention. A request box 1 is displayed to the user, for inputting a number of keywords for a search or query; in the example of figure 1, the inputted search is again /greenhouse effect/. In a way known per se, the search may be limited to part of the database, in the example of figure 1 thanks to line 3. In the example of figure 1, the search is not limited, and the "World Wide Web" selection appears in bold on line 3. "OK" button 5 makes it possible for the user to start the search or query. Current search path is displayed to the user, under the request box; the use of the search path will be explained in reference to figures 2 and 3.

[0033] In response to the search, a number of documents or sites are returned. Retrieval of documents - selection of sites or documents among a database of indexed or partially indexed documents or sites - may be carried out in any way known per se. It is notably possible to use an inverted index such as the AltaVista Search Developer's Kit, sold by AltaVista Company. More specifically, a query inputted in the query box by the user is parsed into an internal representation, which is then translated into a request applied to the inverted index. This request is formulated according to the features supported by the inverted index; usually supported features include ranking, boolean searches, phrase searches, stemming, proximity searches, etc.

[0034] A number of these documents or sites returned by the inverted index are displayed to the user in response to the query. In the example of figure 1, the right-hand part of the lowermost section of the display provides the user with a list 7 of documents and sites. Together with each document, it is possible to display a summary of the document or site, as shown in figure 1 on the line immediately under each document or site. An icon representative of the nature of the answer may also be displayed: in the example of figure 1, icon 9 is representative of a document, while icon 11 indicates that the answer is actually a site. The address for accessing the document may also be displayed, as shown in figure 1 on the line under the summary of each document or site. When a document is part of a given category, this category may be displayed as well. In the example of figure 1, the sixth document in the list 7 of results returned by the query is entitled "Many scientists believe runaway greenhouse effect possible"; this document is actually referenced in the category Climate Change with the access path Society : ...: Environment: Climate Change in the category tree. The eighth document is also referenced in a category. As shown in figure 1, it is possible for the user to select the category displayed under a result. The effects of selecting a category in the list of results are explained in reference to figure 5.

[0035] It should be understood at this point of the description that the directory of categories in this embodiment of the invention is given as an example of predefined or static set of categories. In this specification and in the appended claims, the words "predefined" or "static", as opposed to "dynamic", should be understood in that the set of categories is not amended every time a user inputs a query. Still, it may of course be amended to add or remove categories.

[0036] In addition to supplying a list of documents or sites returned by the query, the invention suggests displaying to the user a series 13 of refinement strategies.

BNSDOCID: <EP_____1182581A1_I_>

These refinement strategies comprise categories selected among the predefined set of categories. In the example of figure 1, categories and keywords are displayed as two separate lists 15 and 17, under the request box and the search path. Thus, the refinement strategies "Society" and "Science" in the example of figure 1 are actually predefined categories.

[0037] The predefined categories may be selected in any process known per se. One may for instance select the categories most frequently associated with the documents or sites returned by the query. In one way or another, a number of categories are selected and displayed to the user, in response to the query.

[0038] As shown in figure 1, one may display several levels of categories. In the example, category "Society" is displayed with several sub-categories, viz. "Issues", "Religion and Spirituality". The number of levels of categories displayed may be selected as discussed above, according to the number of documents or sites returned by the query to be found in a category. The interest of doing so is to guide faster the user into a relevant refinement, by skipping intermediate categories in the hierarchy of categories.

[0039] In addition to selected predefined categories, the refinement strategies displayed to the user comprise a number of dynamically derived keywords. In the example of figure 1, these keywords appear in a list separate from the list of categories. Contrary to this embodiment, refinement strategies could be displayed without identifying predefined categories and dynamically derived keywords. In this case, the user would not distinguish between predefined categories and keywords.

[0040] The keywords depart fundamentally from predefined categories.

[0041] First, categories are limited in number, and are a representation of the world, at the time the database is built. On the contrary, keywords, whatever the way they are constructed, do not constitute a representation of the world, but simply a way to retrieve documents. In a way, they have no meaning per se and are much more low level than categories.

[0042] Second, in view of this difference, the number of categories is much lower than the number of keywords. The number of categories is typically around thousands or tens of thousands; the number of keywords may amount more than the number of plain words in the language. Typically the number of keywords could be several millions or tens of millions.

[0043] Third, categories may not be manipulated, and the mapping of entries to categories does not change when a request is processed. Consequently, when a category is selected, all documents of the category are selected - since the category is per se a set of documents, even if it has a name for the purpose of displaying it to the user. On the opposite, keywords are objects without meaning that can be manipulated while formulating or refining a query. Specifically, when a query is refined by a category, the resulting documents are a subset of the category. When a query is refined by a keyword, the resulting documents need not be associated statically with the keyword.

[0044] The difference between keywords and categories appears in Figures 1-5. Along the search, categories of the higher level disappear, as the user navigates downwards in the hierarchy of categories. This is the case where categories are organised hierarchically, e. g. in a tree or directed acyclic graph.

[0045] In case the invention is applied to a database of textual entries, there is provided an inverted index for retrieving entries. Categories are then necessarily entries of the inverted index, while keywords are not necessarily entries of the inverted index.

[0046] For instance, assume keywords are sequences of words. The keyword "fossil fuels" could then be associated to every document that contains the exact sequence of words "fossil fuels", at the time the database is built. When the keyword "fossil fuels" is selected by the user as a refinement strategy, the query may return documents not only containing "fossil fuels", but also documents containing separate occurrences of the words "fossil" and "fuels". Examples of algorithms for processing keywords - e.g. thanks to stems or synonyms - are given below.

[0047] For dynamically deriving the keywords from the entries returned by the query, one may use any process known per se. In this respect, the keywords may be comprised of words or sequence of words. As displayed in figure 1, it is preferred that the keywords be comprised of sequences of words. One may for instance, in a first phase of operation conducted prior to the queries, derive from all documents in the database a set of relevant sequences of words, e. g. using the algorithm described by Y. Choueka in "Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in Large Textual Databases" (Conference on User-Oriented Content-Based Text and Image Handling, MIT, Cambridge, MA, USA, pp. 609--623, 1988). This step forms a database of keywords. Then, in a second phase of operation conducted in response to the user's query, one may dynamically extract from this keyword database the keywords associated with the set of documents selected by the query and select the most frequently appearing ones as the set of keywords to be displayed to the user,

[0048] For enhancing retrieval effectiveness of the search engine, keywords may also comprise stems (or prefixes), instead of words, or sequences of stems of prefixes. For instance, a stem like "telephon*" covers words like "telephone", "telephones", "telephony" and the like. The stem "mobil*" covers the words "mobile", "mobiles", "mobility", etc. In the example of "mobile telephone", using as a keyword the sequence of stems "mobil* telephon*" provides a more effective use of keywords. In such a setting, a given keyword K is associated to a set S(K) of sequences of words. For instance, for the keyword "mobil* telephon*", the set S("mobil* tel-

BNSDOCID: <EP_____1182581A1_I_>

ephon*") comprise "mobile telephone", but also "mobile telephony", "mobile telephones", "mobility telephoning", and the like. Each word may be automatically associated to its stem, by using the algorithm disclosed in M.F. Porter, An algorithm for suffix stripping, Programs, vol. 14 no. 3, pp. 130-137, July 1980.

[0049] When keywords comprise stems, a document may be associated to a keyword K if the document contains any word or sequence of words in the set S(K) of the keyword K.

[0050] At this point, one should note that the displayed keywords do not necessarily cover all relevant documents or sites, contrary to the suggestion of WO-A-98 49637.

[0051] The invention thus suggests displaying to the user, in response to the query, selected categories as well as dynamically selected keywords. Both may be used for refining the search strategy, as explained now.

[0052] The user may select one or more of the displayed categories and keywords for refining the search strategy. Selecting a displayed category or keyword issues a new query, refined by the newly selected category, or by the newly selected keyword. For instance, selecting the sub-category "Issues " in the example of figure 1 would return the documents and sites concerning greenhouse effect and contained in the predetermined sub-category "Issues" of the category Society. Figure 2 is the corresponding display. It is similar to the display of figure 1, save for the fact that the search path indicates that the search is limited to the category "Issues". In addition, the documents and sites returned by the query are different from the ones of figure 1. Notably, the number of hits or results is 26 instead of 839. Last, figure 2 shows that the keywords and categories are updated, according to the results returned by the query. Limitation through categories, as exemplified on figure 2, normally returns a subset of the set of hits or results provided by the first query.

[0053] By selecting and displaying categories, and allowing a user to refine a query thanks to these categories, the invention makes it possible for the user to benefit from the expertise embodied in the predetermined categories. In addition, by selecting keywords and displaying these keywords, the invention avoids the limitations inherent to any system of predetermined categories. Categories and keywords are provided transparently to the user; navigation among documents thus remains simple and effective, and does not require any specific skill from the user.

[0054] In the embodiment of figure 1, a "minus" sign is provided in front of each of the keywords. In a particular embodiment of the invention, this sign is used for excluding from the query the corresponding keywords. For instance, in the example of figure 1, selecting the "minus" sign in front of "greenhouse gases" may exclude from the query all documents containing the keyword "greenhouse gases". This feature of the invention makes it possible to more easily reduce the number of answers, by excluding those answers containing certain keywords. Selecting the link to a keyword generates a new query, with the said keyword. Figure 3 shows the display obtained by selecting the keywords "fossils fuels" in the example of figure 1. It is again similar to the displays of figures 1 and 2. However, the search path is updated, as well as the list of keywords and categories. The list of results is also different. One should note at this point that the results obtained in such a refinement of the search to a specific keyword are not necessarily a subset of the results of the original query: indeed, this may depend on the algorithms used for the search, and notably of the way sequences of words are handled in this algorithm.

[0055] For instance, assume that the user originally inputs the word /climate/ in the request box. The first query would then return documents containing exactly the word "climate". Assume the user then selects a refinement strategy, e. g. the keyword "fossil fuels" among the list of keywords: the reformulated query may then be /climat* fossil* fuel*/, where "climat*" is the stem for "climate", "fossil*" is the stem for "fossil" and "fuel*" is the stem for "fuels". Contrary to the original query, which only returns documents containing the exact word "climate", the refined query can return documents containing, e.g., "climatic", but not the exact word "climate". In this example, words inputted by the user are replaced by their stems only when the query is refined. The interest of doing so is that replacing words by their stem in user queries - which is a form of query generalisation or query expansion - increases the number of results. If it is carried out at a time the query is not precise enough, the search engine may return too many irrelevant documents; the invention therefore suggests only performing stemming at some point of the refinement process. One may chose to do so after a given number of refinements.

[0056] Note that the same argument applies to refinement by categories; refinement by a category may indeed return more documents than the original query, inasmuch as this query was extended by stemming. However, the refinement by category does provide a subset of the stemmed query.

[0057] In this example, it appears clearly that the documents returned after the query is refined are not necessarily associated to the keyword "fossil fuels". The exact way keywords are handled depends on the inverted index used to retrieve documents.

[0058] It is also possible to allow the user to select more than one category or one keyword as a refinement strategy. Selecting several refinement strategies at the same time may allow the user to restrict the number of documents more easily and faster.

[0059] Figure 4 is a display obtained by refining the search to the keyword "CO2 emissions" in the display of figure 2. The number of results returned is low - 6 in the example -. In this case, it is possible, as exemplified on figure 4, to display the list of documents or sites, with-

BNSDOCID: <EP_____1182581A1_I_>

out any further categories or keywords. This applies notably where the number of results is less than 10, under the assumption that the user may easily browse all the answer, and need not restrict the number of hits.

[0060] In figures 3, 4 and 5, the search path provides links to previous keywords or searches. For instance, in figure 4 or 5, the user may select "Issues" in the search path, and get back to the display of figure 2.

[0061] Figure 5 is another view of the display of the searching tool of figure 1, after the search is limited to a category of the list of results; specifically, the display of figure 5 is obtained when the user selects the category "Climate Change" in the sixth result of the list of results. As explained in reference to figure 1, in response to a user query, documents matching the query may be returned to the user, together with a description and, when available, the categories in which this document is classified. When the user selects on one of these categories in the list, the search engine initiates a new search and displays all documents contained in the category. Preferably, the documents are ranked or sorted according to the request box. In the example of figure 5, category "Climate Change" comprises 122 documents, all of which are displayed in the list of results. The search path shows the path to the category. Subcategories - that is categories referenced in the "Climate Change" category are displayed.

[0062] Selecting a category in the list of results thus allows the user to navigate in all the documents of the category while maintaining the focus on its initial query - in view of the query being displayed in the display box, and possibly used for ranking documents. This capability allows the user to widen its search to documents related to the one that has been found without loosing the focus of its search; in addition, the same navigational tools as before are used, namely, subcategories and keywords. This embodiment may be carried out by using a pure ranking mechanism for queries: no boolean filter is used, except for selecting the category being visited, while navigating in a category.

[0063] Figure 6 is a flowchart of a process according to the invention. In the first step 20, there is provided a database, that is a set of entries - such as documents or sites. Each entry is associated with zero, one or more keywords.

[0064] The database being provided, there is provided at step 22 a set of categories. As discussed above, the set of category is a "static" set. It is advantageous that the categories be organised in a tree structure, for ease of navigation among the categories; however, this is by no means a requirement of the invention. A directed acyclic graph may be used, or even any type of graph. Entries are then mapped to categories.

[0065] At step 24, the user inputs a query - "greenhouse effect" in the embodiment of figure 1.

[0066] At step 26, the results of the query are returned by the search engine. They may be displayed to the user.

[0067] At step 28, the results of the query are analysed for selecting at least one category. The categories are then displayed to the user.

[0068] At step 30, the results of the query are analysed, and keywords are dynamically retrieved. These keywords are then displayed to the user.

[0069] At step 32, selected categories and keywords are displayed to the user, as shown at 13 in figure 1.

[0070] At step 34, the user selects one or more of the displayed categories and keywords. A new query is defined at step 36. The query is applied to the database, and the process may actually go back to step 26. In this manner, new results may be displayed, together with newly selected refinement strategies. Refinement strategies do not only provide restriction of the results obtained in a previous query, but may also provide new results. The search may thus be extended to other more relevant keywords and retrieve more relevant documents or sites.

[0071] The invention may be carried out as follows, in the example of a database of textual entries, using sequences of words or stems as keywords. The entries may comprise documents, or sites.

[0072] Four databases are set up to store respectively the inverted index of the documents words, the keywords associated with the documents, the categories associated with the documents and a summary of each document comprising for instance the document's address on the Internet and its few first lines. All four of these databases may be for instance realised using the AltaVista Search Developer's Kit augmented with a driving program realised in the C language.

[0073] Note that implementers may choose to integrate all four databases into a single database, or that there could also be an inverted index and another database holding both the categories, the keywords, and the summaries, this database being a very simple array of records indexed by the document identifiers returned by the inverted index.

[0074] An indexing phase is first carried out: each of the documents is in turn parsed into a sequence of words; in the case stems are used, each word is associated to its stem by the algorithm described above. Using a hash table data structure, each word/stem pair is then turned into two integer numbers, a word identifier and a stem identifier. Then, assuming that keywords K are specified, e.g., as sequences of stems "S1 ... Sn", as described above, using another hash table, the document is associated to the keywords K such that the sequence "S1 ... Sn", appears in the document. This being done, words, stems, and keywords are used to populate the four databases using the methods described above; the words forming the document are added to the inverted index, together with their stem, the categories possibly associated with the document are added to the category database and to the inverted index, the keywords associated to the document are added to the keyword database (in full or in part), and a summary is built for

BNSDOCID: <EP_____1182581A1_I_>

the document and added to the summary database. Note that the classification of a document into a category is done manually once and for all, whereas the extraction of keywords from the documents is entirely automatic. Also, note that there is no need to add keywords to the inverted index, due to the difference in treatment between categories and keywords, as described above. For instance, assuming the refinement consists in returning only documents that contain the exact keyword "fossil fuels", one may do so by having the inverted index return only entries containing the word "fossil" immediately followed by the word "fuels"; this type of feature is typically provided by modern inverted indexes such as the one suggested above. In this case, "fossil fuels" need not be added to the inverted index as a separate or atomic entry.

[0075] A HTTP server is then set up in order to serve user queries over the Internet. One may for instance use the Apache HTTP server.

[0076] A specialised extension module is added to this HTTP server, in order to carry out the processing associated to the users' queries. This extension module may for instance be realised using the C language and the Apache API, or using the Perl language and the mod_perl Apache module, or the Java language and the JServ Apache module. The specialised module carries out the process of the invention. In answer to a user's query, the module first contacts the inverted index database in order to retrieve a ranked set of documents satisfying the query. The module then contacts all three other databases in order to retrieve the categories, keywords and summaries associated with these documents.

[0077] The module is then able to carry out the keywords and categories selection processes described above; once the relevant categories and keywords are selected, they are combined with the resulting documents for building the HTML answer page to be returned to the user by the HTTP server.

[0078] Figure 7 is a diagram showing a searching tool according to the invention. It shows a computer 40, which represents a user access to the tool. In the example, the tool is accessed through the Internet 42. The searching tool comprises a HTTP server 44; as discussed above, this server receives the queries of the users and returns the answers as HTML documents. The searching tool also comprises a database 46; as discussed above, the database comprises four parts, namely an inverted index 48, a category database 50, a keyword database 52 and a summary database 54. As discussed above, in answer to a user's query, the HTTP server search module first contacts the inverted index database in order to retrieve a ranked set of documents satisfying the query; the documents may be returned simply as a list of documents Ids, as indicated in figure 7. The module then contacts all three other databases in order to retrieve the categories, keywords and summaries associated with these documents; as shown in figure 7, the module may simply send the document ID to retrieve the necessary information.

[0079] Figure 7 further shows a crawler 56 used for referencing web servers 58. The crawler searches for new information available on the Internet, and updates the database.

[0080] The invention was disclosed in the present description in reference to Internet searches, the results of the search being documents and web sites of the World Wide Web. The invention applies more generally to searches among any type of indexed or non-indexed database, provided a number of keywords may be associated to entries of the database. In addition, the entries of the database may be at least partially mapped into categories, for returning categories and allowing the user to refine the search. In this respect, the World Wide Web is a paradigm of a database, while indexed documents or web sites are paradigms of database entries. In the embodiment of the invention disclosed in figures 1 to 5, there is suggested to display a list of entries returned by the queries; the invention may actually be carried out without displaying this list, but simply by displaying refinement strategies to the user.

[0081] Last, the invention is not limited to the description made above. Other ways of populating databases may be used.

Claims

1. A process for searching a database of entries, comprising the steps of:

 - providing a database of entries, at least part of said entries being mapped to a set of categories, at least part of said entries being associated with keywords;
 - in response to a query of a user,

 - selecting categories among said set of categories according to the entries returned by said query;
 - dynamically selecting keywords associated to the entries returned by said query; and
 - displaying to the user said selected categories and said selected keywords.

2. The process of claim 1, wherein the categories are organised in a tree or directed acyclic graph structure.

3. The process of claim 1 or 2, wherein a keyword is a sequence of words or a sequence of stemmed words.

4. The process of claim 1, 2 or 3, wherein the selected categories and selected keywords are displayed

BNSDOCID: <EP_____1182581A1_I_>

similarly.

5. The process of claim 1, 2 or 3, wherein the selected categories are displayed separately from selected keywords.

6. The process of one of claims 1 to 5, further comprising a step of starting a new query when a user activates one of said displayed categories and keywords.

7. The process of claim 6, wherein the step of activating comprises excluding from the said query of the user a displayed category or keyword.

8. The process of claim 6, wherein the step of activating comprises refining the said query of the user to the said category or keyword.

9. The process of one of claims 1 to 8, further comprising displaying to the user a list of entries returned by the query.

10. The process of claim 9, further comprising displaying in said list a category to which at least an entry of said list is mapped.

11. The process of claim 10, further comprising displaying the entries comprised in a category when the user selects said category in said list.

12. The process of claim 11, further comprising ranking the entries comprised in said category before they are displayed.

13. The process of one of claims 1 to 12, wherein categories are hierarchically organised, and wherein the step of displaying comprised displaying categories of different hierarchical levels.

14. A searching tool, comprising a search server (44) for receiving queries from users and transmitting results to users, a database (46) of entries, at least part of said entries being mapped to a set of categories, at least part of said entries being associated with keywords; wherein the search server comprises

- means for searching the database and for selecting categories among said set of categories according to the entries returned by said query;
- means for dynamically selecting keywords associated to the entries returned by said query;

and wherein the results transmitted to the users comprise said selected categories and said selected keywords.

15. The tool of claim 14, wherein the search server is a HTTP server.

16. The tool of claim 14 or 15, wherein the entries are textual entries and wherein the database (46) comprises an inverted index (48), said categories being entries of said inverted index.

BNSDOCID: <EP_____1182581A1_I_>

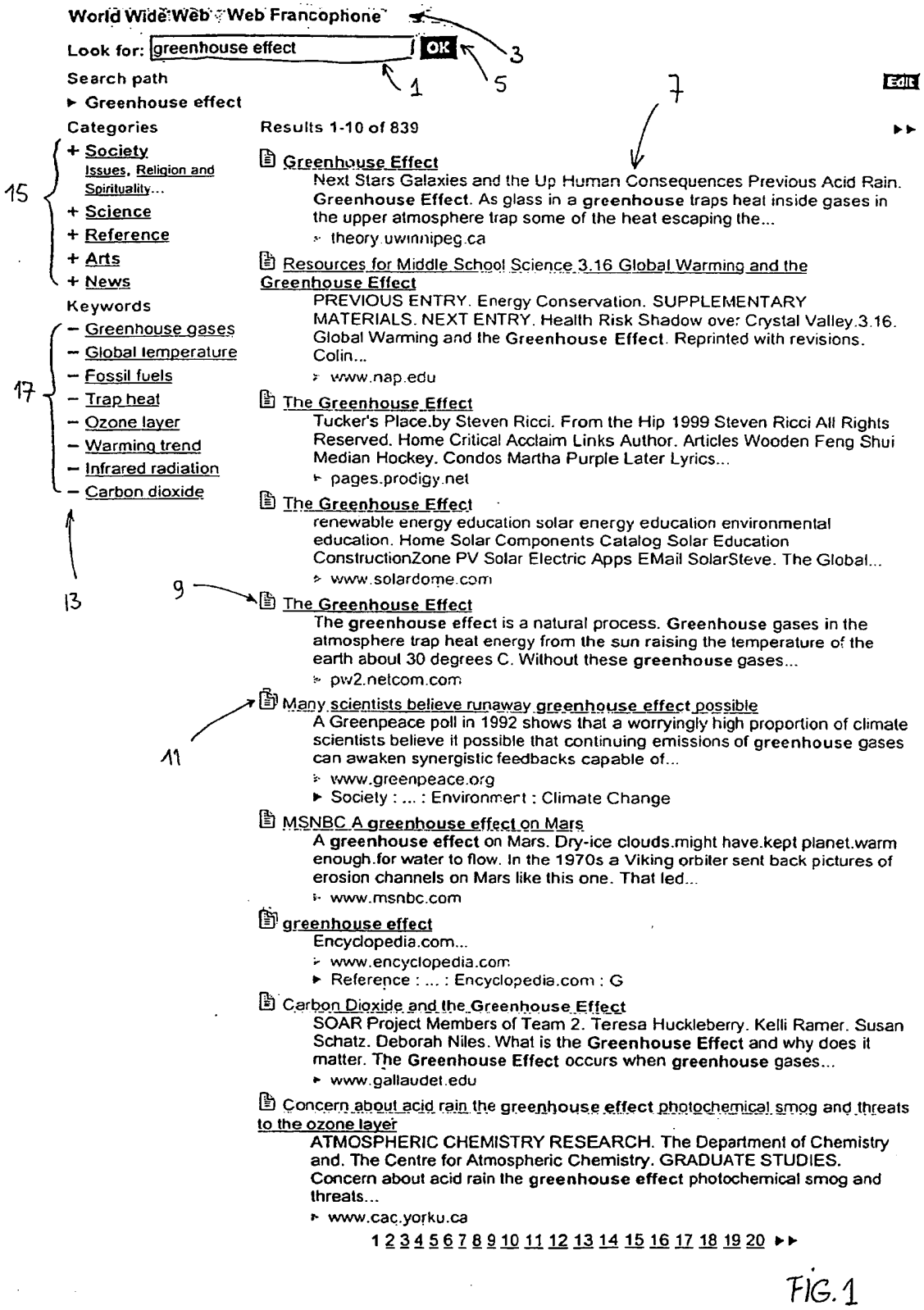


FIG. 1

BNSDOCID: <EP_____1182581A1_I_>

World Wide Web Web Francophone

Look for: greenhouse effect OK

Search path Edit

▸ Greenhouse effect : Issues

Categories

+ Activism

Keywords

- Global warming debate
- Global warming
- Greenhouse gases
- Warming trend
- Greenhouse warming
- Intergovernmental Panel
- Global temperature
- Average global
- Temperature increases
- Global problems
- CO2 emissions
- Lower troposphere
- Mean temperature
- Deep ocean

Results 1-10 of 26 ▸▸

Many scientists believe runaway greenhouse effect possible
A Greenpeace poll in 1992 shows that a worryingly high proportion of climate scientists believe it possible that continuing emissions of greenhouse gases can awaken synergistic feedbacks capable of...
www.greenpeace.org
▸ Society : ... : Environment : Climate Change

Belgium Federal Department of the Environment Climate Change
Climate change web pages of the Belgian federal environment administration. Site has general information on the greenhouse effect policy and legislative information and related links...
www.environment.fgov.be
▸ Science : Environment : Global Change

New Scientist Global Warming Report
Includes a global warming science FAQ global warming politics FAQ a timeline and daily reports from the 1997 Kyoto Conference...
www.newscientist.com
▸ News : ... : Environmental Issues : Climate and Earth Changes

Mainstream Press Misdeeds
Judi Bari's refutation of mainstream press distorting information about Earth First and other activist groups is layed out at the bottom of this editorial page by Jeff Elliott...
www.monitor.net
▸ Society : ... : Free Speech : Subverted

Global Warming Scare Origins
Authoritative account of computer modeling. Leading British scientist explains how an inadequate model started the bandwagon...
www2.prestel.co.uk
▸ Society : ... : Anti-Environmentalism : Global Warming Myths

The Threat of Asteroidal and Cometary Impacts
Report by J. R. Tate of Spaceguard UK. Includes Executive Summary. Japanese research indicates that there is a 1 chance that every major city on the Pacific Rim will receive catastrophic damage from...
dspace.dial.pipex.com
▸ Society : ... : Catastrophes : Large Near Earth Object Impacts

QuoteFile
Politically progressive quotations arranged topically...
www.lancs.ac.uk
▸ Reference : Quotations
▸ Society : ... : Free Speech : Improving

Global Change and Environmental Education Resources
materials and links...
www.gcrio.org
▸ Science : Environment : Education

GCRIO Global Change and Environmental Education Resources
This collection of resources was selected for their relevance to global change and environmental education. Included are a wide range of resources for educators and students at all levels...
gcrio.gcrio.org

Scientists for Global Responsibility
SGR promotes responsible use of science and technology in society and explores the ethical dimension of scientific research ethical controversies global energy and resource issues alternative...
www.sgr.org.uk
▸ Science : Science in Society : Research Ethics
▸ Society : ... : Applied : Science and Technology

1 2 3 ▸▸

FIG. 2

BNSDOCID: <EP_____1182581A1_I_>

World Wide Web Web Francophone

Look for: greenhouse effect OK

Search path Edit

▸ Greenhouse effect : Fossil fuels

Categories

+ Society
+ Science

Keywords

- Greenhouse gases
- Climate system
- Ice caps
- Warming trend
- Global carbon
- General circulation models
- Carbon dioxide
- Global average temperature
- Atmospheric CO2
- Radiative forcing
- Water vapour
- Kyoto conference
- Alternative energy sources

Results 1-10 of 66 ▸▸

Greenhouse Effect
Next Stars Galaxies and the Up Human Consequences Previous Acid Rain. **Greenhouse Effect**. As glass in a **greenhouse** traps heat inside gases in the upper atmosphere trap some of the heat escaping the...
▸ theory.uwinnipeg.ca

The Greenhouse Effect
The **greenhouse effect** is a natural process. **Greenhouse** gases in the atmosphere trap heat energy from the sun raising the temperature of the earth about 30 degrees C. Without these **greenhouse** gases...
▸ pw2.netcom.com

A Primer on Ozone Depletion The Environmental Citizenship Series
The Green Lane Ozone Home Page Search Français Back to Ozone Depletion. A Primer on Ozone Depletion. The Environmental Citizenship Series. ACKNOWLEDGEMENTS. We would like to take this opportunity to...
▸ www.ec.gc.ca

Landover Baptist Church
If religion is the drug of the masses then this is the church that gives you the munchies. The kind of stuff that got Jerry Falwell kicked out of Sunday School. Naughty hilarious and almost...
▸ www.landoverbaptist.org
▸ Society : Religion and Spirituality : Humor

New Scientist Global Warming Report
Includes a global warming science FAQ global warming politics FAQ a timeline and daily reports from the 1997 Kyoto Conference...
▸ www.newscientist.com
▸ News : ... : Environmental Issues : Climate and Earth Changes
▸ Science : Environment : Global Change
▸ Society : ... : Environment : Climate Change

The EPA Global Warming Site Glossary of Climate Change Terms
A B C D E F G H I J K L M N O P R S T U V W Y Z. Absorption of Radiation. The uptake of radiation by a solid body liquid or gas. The absorbed energy may be transferred or re-emitted. Acid Rain. Also...
▸ www.epa.gov

AGO What's New
What's New. May 2000. Week of 8 12 May. International. Notice of the Australian Government Review of The Intergovernmental Panel on Climate Change Third Assessment Draft Reports and invitation to...
▸ www.greenhouse.gov.au

Resources for Middle School Science 5.8 Matter and Energy in the Biosphere
PREVIOUS ENTRY. The Local Environment. CORE MATERIALS. NEXT ENTRY. Middle Grades Science A Problem-Solving Approach Sixth Grade.5.8. Matter and Energy in the Biosphere.2nd ed. Francis M. Pottenger...
▸ www.nap.edu

THE LAB Big Fat Science
by Bernadette Hobbs. For 15 years now we've been pelted with details about the **greenhouse effect** and global climate change. But does a worldwide obsession with composting and rifling through garbage...
▸ www.abc.net.au

A Climate-Protection Technology But Not a Magic Bullet The Global Citizen 2000 02 24
Donella Meadows The Global Citizen February 24 2000. A Climate-Protection Technology But Not a Magic Bullet. The February issue of Scientific American tells of a new technology that makes me both...
▸ iisd.ca

1 2 3 4 5 6 7 ▸▸

FIG.3

BNSDOCID: <EP_____1182581A1_I_>

World Wide Web Web Francophone

Look for: greenhouse effect OK

Search path Edit

▸ Greenhouse effect : Issues : CO2 emissions

Results 1-6 of 6

Global Warming Scare Origins
Authoritative account of computer modeling. Leading British scientist explains how an inadequate model started the bandwagon...
▸ www2.prestel.co.uk
▸ Society : ... : Anti-Environmentalism : Global Warming Myths

Still Waiting For Greenhouse
Non-believers in global warming present evidence to show that global warming has been grossly exaggerated...
▸ www.vision.net.au
▸ Society : ... : Environment : Climate Change

FMN Policy Spotlight Global Warming
resources casting a skeptical eye at sometimes outlandish environmental claims...
▸ www.free-market.net

Population Control How Many Are Too Many
Article from IMPACT Press a bimonthly socio political magazine based in Orlando FL...
▸ www.impactpress.com
▸ Society : ... : Population : Pro-population control

Global Warming Projections for the New Millennium
Environmental Defense Fund EDF charts vividly demonstrate what may be in store for Earth unless greenhouse-gas pollution is brought under control...
▸ www.edf.org
▸ Society : ... : Activism : Global Warming

Environmental Effects of Increased Atmospheric Carbon Dioxide Anti Global Warming Petition Project
The Anti Global Warming Petition Project published a scientific-appearing article contradicting many of the IPCC findings. Much of the information is sound even if the conclusions are biased...
▸ sitewave.net

FIG 4.

BNSDOCID: <EP_____1182581A1_I_>

Look for: greenhouse effect OK

Search path Edit

▸ Society : Issues : Environment : Climate Change

Categories

+ Activism
+ News
+ Proposed Tech Fixes

Results 1-8 of 122 ▸▸

Many scientists believe runaway greenhouse effect possible
A Greenpeace poli in 1992 shows that a worryingly high proportion of climate scientists believe it possible that continuing emissions of **greenhouse** gases can awaken synergistic feedbacks capable of generating a runaway **greenhouse effect**. .
▸ www.greenpeace.org

Global Warming Understanding Greenhouse Gases for Education and Learning
Learn the Science of Global Warming. . Global Warming Understanding **Greenhouse** Gases for Education and Learning. The Science of Global Warming. (...) on experimental tool for investigating global warming and the **greenhouse effect**. Several background topics are included irradiation photons molecules and ...
▸ seeds2lrn.com

Still Waiting For Greenhouse
...that global warming has been grossly exaggerated. . Still Waiting For **Greenhouse**. This web site is best viewed with a resolution (...) Tasmania.25 July 1999. John L. Daly. *Still Waiting for* **Greenhouse**. A Lukewarm View of. Global Warming.from Tasmania.by. ...
▸ www.vision.net.au

Belgium Federal Department of the Environment Climate Change
Climate change web pages of the Belgian federal environment administration. Site has general information on the **greenhouse effect** policy and legislative information and related links. .
▸ www.environment.fgov.be
▸ Science : Environment : Global Change

Belgium Federal Department of the Environment Climate Change
Climate change web pages of the Belgian federal environment administration. Site has general information on the **greenhouse effect** policy and legislative information and related links. .
▸ www.environment.fgov.be

FIG 5

Environmental Effects of Increased Atmospheric Carbon Dioxide Anti Global Warming Petition Project
...markedly increased plant growth rates. Predictions of harmful climatic **effects** due to future increases in minor **greenhouse** gases like CO2 (...) been responsible for much of this increase. But the **effect** on the environment is likely to be benign. **Greenhouse** gases ...
▸ sitewave.net
▸ Society : ... : Activism : Global Warming

Global Warming Projections for the New Millennium
...vividly demonstrate what may be in store for Earth unless **greenhouse**-gas pollution is brought under control. . Environmental Defense Global (...) may be underestimated in that case. In each case **effects** are felt far into the future because **greenhouse** gases continue ...
▸ www.edf.org

The Center for Energy Climate Solutions CECS
...Climate Solutions CECS a one-stop shop that helps reduce **greenhouse** gas emissions with practical advice tools and technologies provided (...) Joseph Romm. How to Boost Profits and Productivity by Cutting **Greenhouse** Gas Emissions. Welcome to Cool-Companies.org the official ...
▸ www.cool-companies.org

1 2 3 4 5 6 7 8 9 10 11 12 13 ▸▸

BNSDOCID: <EP_____1182581A1_I_>

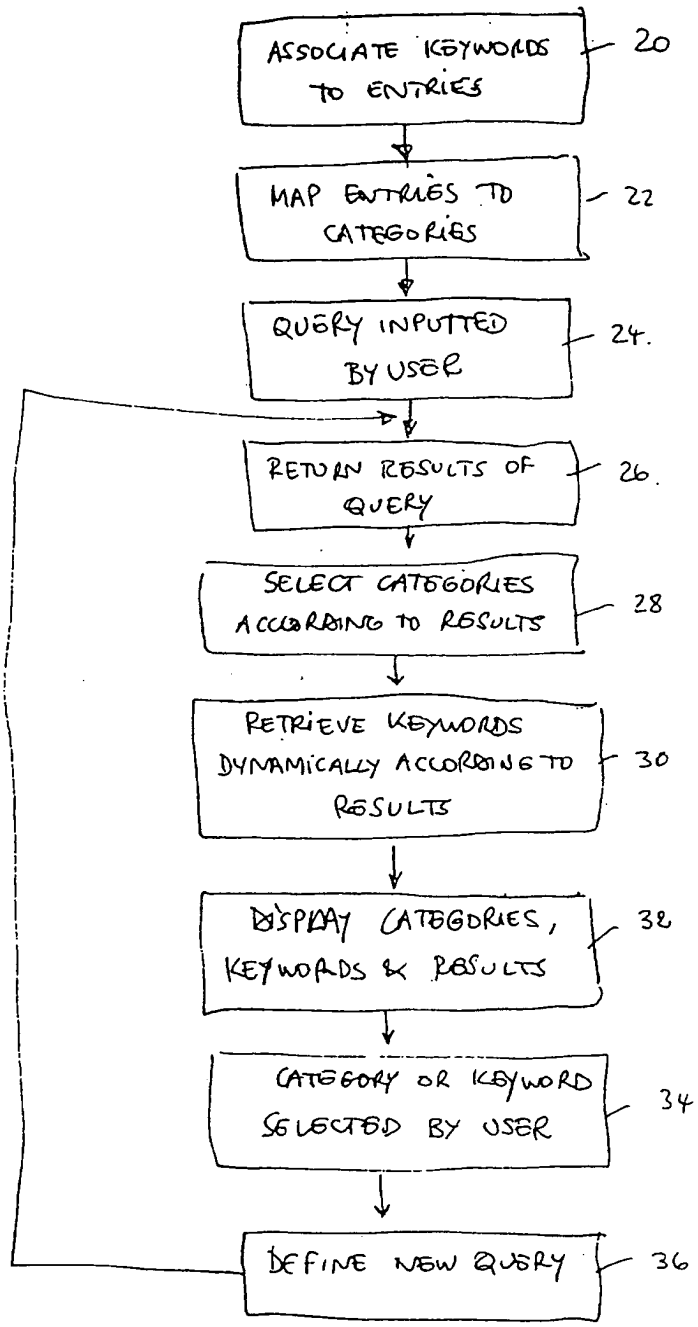


FIG 6.

BNSDOCID: <EP_____1182581A1_I_>

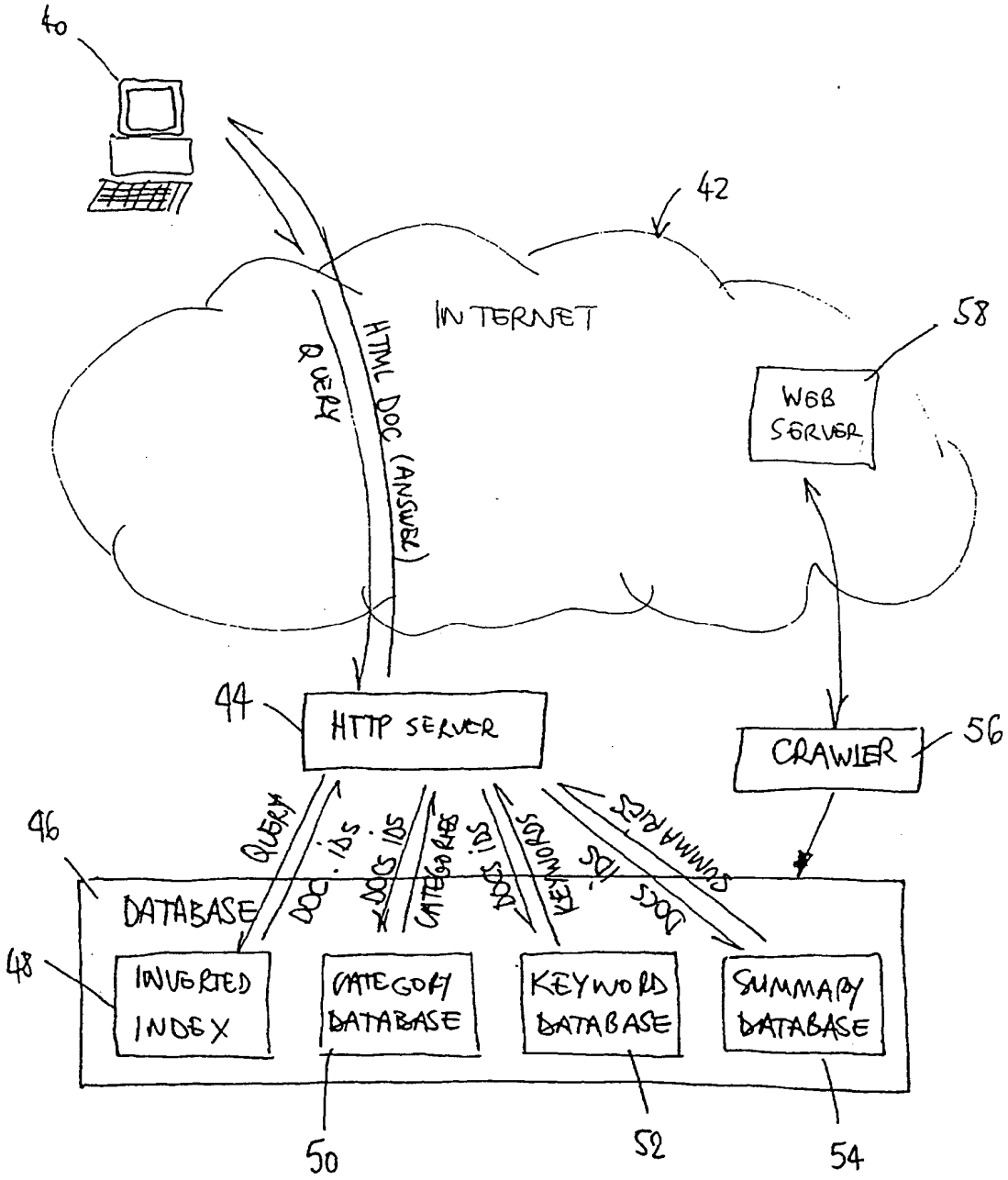


FIG 7.

BNSDOCID: <EP_____1182581A1_I_>

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 00 40 2311

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
X	WEISS R ET AL: "HYPURSUIT: A HIERARCHICAL NETWORK SEARCH ENGINE THAT EXPLOITS CONTENT-LINK HYPERTEXT CLUSTERING" HYPERTEXT '96. 7TH. ACM CONFERENCE ON HYPERTEXT. WASHINGTON, MAR. 16 - 20, 1996, ACM CONFERENCE ON HYPERTEXT, NEW YORK, ACM, US, vol. CONF. 7, 16 March 1996 (1996-03-16), pages 180-193, XP000610424 ISBN: 0-89791-778-2 * the whole document * ---	1-16	G06F17/30
X	ANONYMOUS: "Taxonomized Web Search" IBM TECHNICAL DISCLOSURE BULLETIN, IBM CORP. NEW YORK, US, vol. 40, no. 5, 1 May 1997 (1997-05-01), pages 195-196, XP002133594 ISSN: 0018-8689 * the whole document * ---	1-4,9-16	
X	EGUCHI K ET AL: "Adaptive and incremental query expansion for cluster-based browsing" PROCEEDINGS. 6TH INTERNATIONAL CONFERENCE ON ADVANCED SYSTEMS FOR ADVANCED APPLICATIONS. PROCEEDINGS. 6TH INTERNATIONAL CONFERENCE ON DATABASE SYSTEMS FOR ADVANCED APPLICATIONS, HSINCHU, TAIWAN, 19-21 APRIL 1999, pages 25-34, XP002175471 1999, Los Alamitos, CA, USA, IEEE Comput. Soc, USA ISBN: 0-7695-0084-6 * the whole document * --- -/--	1,3-12, 14-16	TECHNICAL FIELDS SEARCHED (Int.Cl.7) G06F

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	21 August 2001	Abbing, R

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

BNSDOCID: <EP_____1182581A1_I_>

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 00 40 2311

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
A	HEARST M A ET AL: "CAT-A-CONE: AN INTERACTIVE INTERFACE FOR SPECIFYING SEARCHED AND VIEWING RETRIEVAL RESULTS USING A LARGE CATEGORY HIERARCHY" PROCEEDINGS OF THE 20TH ANNUAL INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL. PHILADELPHIA, PA, JULY 27 - 31, 1997, ANNUAL INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEV, 27 July 1997 (1997-07-27), pages 246-255, XP000782010 ISBN: 0-89791-836-3 * the whole document * ---	1-16	
A	WO 98 09229 A (TELEVITESSE SYSTEMS INC ;STREATCH PAUL (CA); REED JIM (CA)) 5 March 1998 (1998-03-05) * the whole document * ---	1-16	
A	US 5 933 827 A (COLE GARY LEE ET AL) 3 August 1999 (1999-08-03) * abstract * * column 2, line 36 - column 2, line 54 * * column 4, line 18 - column 6, line 22 * * column 10, line 34 - column 11, line 18 * -----	1-16	TECHNICAL FIELDS SEARCHED (Int.Cl.7)

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	21 August 2001	Abbing, R

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

BNSDOCID: <EP_____1182581A1_I_>

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 00 40 2311

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

21-08-2001

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9809229 A	05-03-1998	CA 2184518 A	01-03-1998
		AU 4007497 A	19-03-1998
		EP 0922260 A	16-06-1999
US 5933827 A	03-08-1999	NONE	

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82

This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☑ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☐ FADED TEXT OR DRAWING
- ☐ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.
As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.